Global Governance Institute

GGI Report 1/2025

Towards a Global Agency for Governing Artificial General Intelligence

Challenges and Prospects

Justin B. Bullock

www.globalgovernance.eu

About The Global Governance Institute

The Global Governance Institute (GGI) is an independent, non-profit think tank based in Brussels. GGI brings together policy-makers, scholars and practitioners from the world's leading institutions in order to devise, strengthen and improve forward-looking approaches to global governance.

Our mission is to promote comprehensive research, cutting-edge analysis and innovative advice on core policy issues, informed by a truly global perspective. This also includes raising awareness about major challenges of global governance among the general public.

Our vision is a more equitable, peaceful and sustainable global order based on effective but accountable international organizations, the global rule of law and the empowerment of the individual across borders and cultures. GGI places particular emphasis on the improvement of the United Nations system and its mutual reinforcement with strong regional organisations.

About the Author

Justin Bullock is a Non-Resident Senior Fellow in the AI and Global Governance Programme. He is also an Associate Professor Affiliate of Governance at the University of Washington in the Evans School of Public Policy and Governance and a world-renowned scholar in Public Policy, Public Administration, Governance, and Artificial Intelligence. Dr. Bullock is also a Senior Researcher at Convergence Analysis where he leads the research of Project AI Clarity.

How to cite:

Justin Bullock (2024) Towards a Global Agency for Governing Artificial General Intelligence: Challenges and Prospects, GGI Report 2024, Brussels: Global Governance Institute

GGI Reports

Published on 14 January 2025

The Global Governance Institute (GGI) ASBL is a non-profit foundation registered under Belgian law. Registration No. 831.178.152

Contents

Abstract **4**

Introduction **5**

Part 1: AGI Types **7**

Narrow AI & General AI 8

General AI: Benefits & Hazards 8

Unitary Agent AGI 10

Open Agency AGI 10

A Governance Argument for Open Agency AGI 11

Part 2: AGI Governing Institutions **12**

AI Governance & AGI Governance 13

AI Governance of Narrow AI Systems 13

AI Governance of General AI Systems 13

AI Governance of AGI 15

Global AGI Agency Governance Components 16

Part 3: A Global AGI Agency Proposal **18**

Introduction to the Proposal 18

Core Elements of the Global AGI Agency 19

1. Institutional Framework 19

2. Global Collaboration and Integration 21

3. Open Agency Model & Structured Transparency 22

4. Democratic Accountability and Access 25

Proposal Summary 26

Part 4: Challenges **27**

1. International Cooperation Challenges 27

2. Centralization of Power Risks 27

3. Innovation and Competition Concerns 28

4. Private Sector Resistance 28

5. Representation and Fairness Issues 28

6. Complexity of Governance 29

7. Technical Challenges of the Open Agency AGI Model 29

8. Democratic Accountability Limitations 29

9. Security and Information Risks 30

10. Adaptability and Future-Proofing Concerns 30

Part 5: Conclusion **31**

A Global AGI Agency Proposal

Justin B. Bullock[3] [4]

Abstract

This paper argues that there are significant plausible benefits to creating an Open Agency type Artificial General Intelligence (AGI) and having a Global AGI Agency as its primary controller, but that there are also immense challenges to these significant benefits. After a brief exploration of the history of digital computation and global governance, the paper begins with an examination of two AGI types: Unitary Agent and Open Agency. Then, it provides a selected literature review of approaches to AI and AGI governance. The AGI governance literature is sparse, but a few AGI governance proposals are discussed. Next, it examines a proposed scenario in which the world has created a single AGI and that AGI is governed, in large part, by a Global AGI Agency. The Global AGI Agency which has four proposed core elements: (1) an institutional framework involving joint support from the UN and IEEE and from a coalition of national governments and private sector companies, (2) global collaboration and integration across key AI powers and companies, (3) the Open Agency AGI model itself, organized around principles of structured transparency, and (4) mechanisms for democratic accountability and access. Following the discussion of the Global AGI Agency proposal, the paper describes 10 challenges for this proposal. These 10 challenges include: international cooperation, centralization of power, innovation and competition concerns, private sector resistance, representation and fairness, complexity of governance, technical challenges of the Open Agency AGI model, democratic accountability limitations, security and information risks, and adaptability and future-proofing concerns. In the conclusion the author reconsiders the key points from the report and reminds the reader of the important challenge of good AGI governance.

[3] Global Governance Institute, Convergence Analysis, Texas A&M University

[4] The author would like to thank the Future of Life Institute for supporting this work

Introduction

The pace at which technology reshapes our world has been gaining steam since the industrial revolution. This great reshaping has caused significant changes to our social, economic, and environmental ecosystems. Steam powered ships, continent traversing trains and radio waves, and global electrification of homes and factories have impacted how we interact with one another and our environment. These technologies also changed how governments exert power and influence over one another and the individuals that they govern. By the mid-20th century the manipulation of the atom had created atomic bombs, which were birthed with help of digital computation, and this birth was accompanied by death, destruction, and an end to WWII.

In the wake of WWII, both atomic bombs and digital computation improved in their technical capacities, while continuing the trend of technological capabilities that reshaped sociological, cultural, and environmental ecosystems. Given the horrors of WWII, their technological requirements, and their violent nature, atom bombs remained accessible only to a very few well-resourced and motivated nation states throughout the remainder of the 20th century and until present day. And while there was a buildup of the number of bombs and of their destructive capacity, the global stockpile of nuclear weapons peaked around 1988[5], and the destructive capacity of a single bomb peaked in 1961 with the detonation of the Tsara bomba[6].

Digital computation has taken a different route following its own birth with the ENIAC in 1945[7]. While digital computers began as the property of nation states and then academic research labs, by 1959 the advance of semiconductor technology[8] began to pave the way for personal, at home, use for anyone who could afford the machine. Not long after this, digital computation was beginning to be networked across individual machines, giving birth to a new technology in the internet. Networked digital computing, and continued progress in digital computation capacity issued in the digital revolution, which further reshaped our various global ecosystems. Networked digital computation now spans the entire globe, reshaping so many aspects of our world. Global electrification provided the backbone for the global spread of digital computation. Where there is electricity, now, there is very likely to be digital computation. And, unlike atom bombs, networked digital computation powered new industries, wealth creation, and countless improvements in services and products across the globe. Open-sourcing and shared technical standards allowed the benefits of vast increases in computational capacity to be placed in the heads of a majority of individuals in

[5] https://fas.org/initiative/status-world-nuclear-forces/

[6] https://www.britannica.com/topic/Tsar-Bomba

[7] https://en.wikipedia.org/wiki/ENIAC#Role_in_the_hydrogen_bomb

[8] https://en.wikipedia.org/wiki/Personal_computer

the world, with an estimated 68% of the world's population being internet users in 2023[9] and over 94% of Americans having access to the internet in 2024[10].

The powerful capacities of atom bombs and networked digital computation brought twin challenges for the world: 1) destructive capacity of individual states of the like never before seen and 2) global interconnectedness culturally, economically, socially, and politically. These destructive state capacities have brought about an expansion in governance focus from the nation to the globe. As a consequence, as early as the League of Nations following WWI, the UN following WWII, and global standard setting technical bodies growing throughout the 20th century, various tools have developed for ensuring that the most powerful actors in the world have some forms of oversight, accountability, and pathways for coordination. These global governance mechanisms remain imperfect, but provide important measures of both protection and access to the peoples of the world. Protection from harm and access to opportunities from beyond the borders of the particular state in which they live.

It is into this world that Artificial General Intelligence (AGI) is currently being birthed. AI's journey has its origins alongside the development of electrification and digital computation. Electricity powers the ability of machines to perform computation, and improvements in digital computation capacity have powered advances in AI. The progress of AI has been uneven but steady since the 1950s. Early systems of AI were knowledge and manipulation systems that gained some basic narrow skills. Over time, as more computation became available, more computation-intensive AI-architectures began to emerge. However, it was not until very large amounts of computation were married with very large amounts of data, made accessible thanks to the success of the internet, that the field of AI began to approach its goal of human-level capabilities. While there were many successes in the various areas of narrow AI, it is now, on the cusp of AGI, that frontier AI systems are beginning to exhibit very powerful capabilities.

Advanced frontier AI systems are already superhuman in specific capability areas, and there are now a multitude of very well-resourced actors that are actively seeking to build AGI systems that are as capable as any human on essentially any task. As will be discussed in more detail in a later section, there are multiple plausible pathways to creating AGI, and these different pathways create different opportunities and risks to humanity. In short, it matters what type of AGI is created. For much of the field's history researchers have focused on creating unitary agents that have general human level capabilities. However, researchers are now exploring an alternative approach: an Open Agency AGI. Both of these types of AGI could

[9] https://web.archive.org/web/20240319110853/https://www.internetworldstats.com/stats.htm

[10] Internet Usage Statistics In 2024 – Forbes Home

provide power to individual actors on a scale not seen since the development of the atom bomb, requiring global governance strategies to carefully navigate the immense power these systems present to humanity.

Global governance strategies themselves have evolved since the failed League of Nations and the mixed record of success of the United Nations. Organizations like the International Atomic Energy Agency (IAEA), the Financial Action Task Force (FATF), and the International Civil Aviation Organization (ICAO) are intergovernmental, multi stakeholder organizations that globally bring together stakeholders, form norms and standards, and perform monitoring and enforcement roles to different industries. Organizations like the North Atlantic Treaty Organization (NATO) bring together powerful nation states to help coordinate the use of powerful weapons. The paper "International Institutions for Advanced AI[11]" also identifies numerous in progress attempts at applying global governance to AI systems including: Organisation for Economic Co-operation and Development (OECD), the Global Partnership on AI (GPAI), the Group of 7 (G7) Hiroshima Process, the International Telecommunication Union (ITU), as well as by private sector initiatives like the Partnership on AI, the ML Commons, and International Standards Organization (ISO) and International Electrotechnical Commission (IEC) standard-setting initiatives. However, as the same paper notes, these efforts are likely to be insufficient for advanced AI, including AGI.

It is from these general points that the rest of the paper builds. In Part 1, it more fully describes the two major AGI types, discusses tradeoffs across the benefits and hazards of each, and then provides an argument for the benefits of creating an Open Agency AGI instead of a Unitary Agent AGI. Following this, in Part 2 the paper provides a more detailed description of current approaches to international governance of AI and frontier AI, describes the more promising governance approaches to AGI, and selects elements from key proposals to serve as the basis for a Global AGI Agency. In Part 3 I provide a high-level description for four core elements of the Global AGI Agency Proposal. In Part 4 the paper provides a discussion of the drawbacks and limitations of this proposal. And, finally, it offers some concluding thoughts in the conclusion.

Part 1: AGI Types

There are different *types* of AGI, and it matters what sort of AGI is developed. Additionally, the risks that any general AI system presents is different than for a narrow AI system. Most of the AI Governance schemes proposed thus far, are directed at narrow AI systems that currently exist, with some starting to explore more general systems for frontier AI models. This section briefly notes a few key differences between narrow AI and general AI. Then it lays

[11] https://arxiv.org/pdf/2307.04699

out some of the key general benefits and hazards of general AI. Next, it discusses the key differences between general AI systems and AGI in particular and the two main *types* of AGI discussed in the literature: Unitary Agent and Open Agency types. Finally, it briefly makes the case for the Open Agency type from a governance perspective.

Narrow AI & General AI

AI systems have usefully been described as either being narrow AI or general AI. While there is no definitive demarcation that separates these two categories, narrow AI is generally understood as a system that can complete one to a few different types of tasks and general AI is understood as a system that can complete many different types of tasks. Generally speaking, historically, AI systems have been narrow AI systems. That is, they may be good at playing a specific game or making financial trades or translating across different languages. In this way narrow AI is the extension of automation to new classes of digitalized tasks. And, as a consequence of this, single AI systems have become as good or better than humans at these specific tasks. However, narrow AI, by definition is limited, it is applied to some small subset of overall tasks that a human might complete.

While narrow AI is limited, this does not mean that it is insignificant. In fact, narrow AI has already dramatically reshaped our social, economic, and cultural worlds by “simply” tailoring and distributing information in new ways, at scale, and quickly. Three clear examples are: 1) ads delivered from search engines and on social media, 2) algorithms for social media feeds, and 3) transportation routing. Notice that these seem like fairly narrow tasks for an AI to be completing, while at the same time, recent history also demonstrates that very competent narrow AI can transform entire industries such as advertising, marketing, news, transportation, and shopping. Taken together, narrow AI systems have already significantly reshaped the global economy and the structure of numerous industries.

With these capabilities and various impacts already in place, general AI systems have recently been created and deployed. These systems are referred to variously as large language models (LLMs), large multimodal models (LMMs), and frontier AI models. These general AI systems can be applied directly to a wide array of tasks and can make use of various tools to complete other tasks. These are the first generation of general AI systems, and these general AI systems have a different set of benefits and hazards as compared to narrow AI systems.

General AI: Benefits & Hazards

General AI systems have the capacity to complete a wide set of tasks that are typically completed by humans, and to complete these tasks at greater scale and improved efficiency. The possible benefits of the deployment of such a technology are numerous. However, it is also hard to predict the specific benefits and harms of such systems. As with the deployment of narrow AI systems, consequences arise and propagate throughout the various human ecosystems, even when the task being automated is a fairly narrow set of tasks.

Once general AI systems have the technical capacity to complete general sets of tasks, they are also likely to be able to complete these tasks more quickly than humans and at a large scale. This suggests that these more general systems will likely have even larger impacts on human ecosystems that narrow AI systems. For example, general AI systems could greatly expand access to quality education, healthcare, electricity, emergency response, food, housing, and entertainment throughout the world. Furthermore, general AI systems could plausibly complete essentially all dangerous or dirty tasks that most humans do not want to complete or are hazardous to complete. It may furthermore be the case that capital is put towards even more productive ends generating additional vast amounts of profit and increases in overall productive output. This plausibly points to a world with a greatly increased Global World Product and great expansions in both the necessities and the basic leisures of life.

While general AI systems provide the possibility of immense benefits, as a general purpose system they also provide the capabilities to cause immense harm. These harms have been classified in numerous ways in the literature, however this paper focuses on two general types of hazards from general AI systems. The first are misuse risks. This is when general AI systems are used deliberately by some other actor to cause harm. The second are agentic risks. These are hazards that arise from giving the general AI systems autonomy to make and execute decisions. Given general AI systems can complete wide sets of tasks at a level comparable to humans and at larger scale and more quickly, then the deliberate misuse of these systems could also be conducted on a wide set of tasks, quickly, and at scale. Additionally, if the general AI system itself is given the capacity to make and execute decisions, then it becomes an actor or an AI agent that, if not properly goal-aligned or process-aligned, could cause immense harm.

Given the plausible immense benefits and immense harm general AI systems present to the world, it is worth carefully considering the *type* of general AI systems that humans might create. It may have been noticed by an astute observer that the common term AGI has been avoided in the discussion of narrow and general AI. This is purposeful. So far we've explored narrow AI, which we've defined as AI systems that can perform some small set of tasks well and general AI, which we've defined as AI systems that can perform some wide set of tasks well. The term AGI typically implies an AI system can complete all or almost all tasks that a human could complete. With these definitions in mind, it is clear that we already have both narrow AI and general AI systems, but that the AGI bar has not quite been met.

For the purposes of this report, a world in which only one AGI system exists is being assumed. Thus it seems that the form of that AGI is likely to be important to that world. To this end, and building directly from previous work, two types of possible AGI are briefly explored: 1) Unitary Agent AGI & 2) Open Agency AGI.

Unitary Agent AGI

The deep learning paradigm mapped with the strategy of reinforcement learning agents has been the dominant strategy in creating AI agents that can complete increasingly general tasks. The idea is that there is one unitary agent that takes inputs, analyzes them, and produces outputs and actions. The Unitary Agent AGI is nicely described in Eric Drexler's "The Open Agency Model[12]" as a contrast to the Open Agency type of AGI. Drexler describes the Unitary Agent type as follows:

"The unitary-agent model typically carries assumptions regarding goals, plans, actions, and control.

Goals: *Internal to an agent, by default including power-seeking goals*

Plans: *Internal to an agent, possibly uninterpretable and in effect secret*

Actions: *Performed by the agent, possibly intended to overcome opposition*

Control: *Humans confront a powerful, potentially deceptive agent*

The typical unitary-agent threat model contemplates the emergence of a dominant, catastrophically misaligned agent, and safety models implicitly or explicitly call for deploying a dominant agent (or an equivalent collective system) that is both aligned and powerful enough to suppress unaligned competitors everywhere in the world."

The Unitary Agent type suggests a tightly coupled process within the agent's own reasoning about its goals, plans, and actions. As Drexler, and many others have noted, this type of AGI may be power-seeking, difficult to interpret, and may take actions to ensure its survival, all possibly leading to a deceptive agent that is more powerful than any single human and plausibly most collections of humans.

Frontier AI models, including Large Language Models and Large Multi-Modal Models, as the most general AI systems to date, also point at a way to construct a different type of AGI, the Open Agency type.

Open Agency AGI

The Open Agency type of AGI is proposed by Drexler in The Open Agency Model[13] as well. As compared to the Unitary Agent types, Drexler describes the Open Agency type as follows:

"Trained on prediction tasks, LLMs learn world models that include agent

[12] https://www.lesswrong.com/posts/5hApNw5f7uG8RXxGS/the-open-agency-model

[13] https://www.lesswrong.com/posts/5hApNw5f7uG8RXxGS/the-open-agency-model

behaviors, and generative models that are similar in kind can be informed by better world models and produce better plans. There is no need to assume LLM-like implementations: The key point is that generation of diverse plans is by nature a task for generative models, and that in routine operation, most outputs are discarded.

These considerations suggest an "open-agency frame" in which prompt-driven generative models produce diverse proposals, diverse critics help select proposals, and diverse agents implement proposed actions to accomplish tasks (with schedules, budgets, accountability mechanisms, and so forth).

Goals, plans, actions, and control look different in the open-agency model:

Goals: *Are provided as prompts to diverse generative models, yielding diverse plans on request*

Plans: *Are selected with the aid of diverse, independent comparison and evaluation mechanisms*

Actions: *Incremental actions are performed by diverse task-oriented agents*

Control: *Diverse, independent monitoring and evaluation mechanisms guide revision of plans"*

The open agency type of AGI looks more like an organization or institution or *agency* than an *individual*. In the open agency type, numerous, more narrow, bounded agents with the agency complete tasks. Tasks such as goals and plans are separated, distinct tasks that generate their own outputs to be observed, inspected, and incorporated into the rest of the decision making process. Actions are then taken by various action-oriented agents that differ from agents focused on plans and interpreting goals given as prompts. In the open agency model, humans provide prompts for generating proposals, use their preferences to evaluate those proposals, oversee the implementation of those proposals by task agents and service provision, and finally review reports about the actions taken by the agency and the impacts those actions had.

A Governance Argument for Open Agency AGI

With these two types of AGI described by Drexler, he goes on to highlight how these two types also provide partial reframings for classic AI safety concerns. Drexler notes:

"Basic challenges in AI safety — corrigibility, interpretability, power seeking, and alignment — look different in the agent and agency frames:

Corrigibility:

Agents: Goal-driven agents may defend goals against change.

Agencies: Generative planning models respond to goals as prompts.

Interpretability:

Agents: Plan descriptions may be internal and opaque.

Agencies: Plan descriptions are externalized and interpretable.

Power seeking:

Agents: Open-ended goals may motivate secret plans to gain power.

Agencies: Bounded tasks include time and budget constraints.

Alignment:

Agents: Humans may have only one chance to set the goals of a dominant agent.

Agencies: Humans engage in ongoing development and direction of diverse systems."

As Drexler highlights, the Open Agency type has some benefits relative to the Unitary Agent type. Agencies can evaluate various possible goals and receive feedback both from humans and bounded AI agents throughout this process, providing for improvements in corrigibility. These goals can then lead to the creation of various sets of plans that also receive feedback from both humans and bounded AI agents, improving interpretability of AI plans. In the agency setting there are bounded agents performing bounded tasks rather than general agents pursuing open ended goals, this helps mitigate power seeking behavior. And finally, the agency type more clearly allows for an ongoing development, improvement, and shifting of behavior by human feedback, which should aid in improving alignment of the AGI with humans.

In addition to the reframing of AI safety concerns, the Open Agency type appears to have the same general sets of advantages as the Unitary Agent type. That is both types of systems are assumed to have the capability of completing most if not all tasks completed by humans. In this way, they are likely to have similar benefits from a capabilities approach. Additionally, the Open Agency type allows for more clear inspection of inputs, throughputs, and outputs of the system. As a consequence of this, an additional governance benefit to the Open Agency type is that it should be more transparent, interpretable, and allow for clear lines of accountability.

Taken together, in a world with one AGI, it seems more desirable for that system to be the Open Agency AGI type.

Part 3 of this report will describe how an Open Agency type AGI can form the basis for a Global AGI Agency, which is meant to govern the Open Agency type AGI. However, before turning to this specific proposal, let us review some relevant aspects of AI Governance.

Part 2: AGI Governing Institutions

For Part 2 of this report, we will shift from examining *types* of AGI to the *institutions* that will govern their use. An overview is offered of the difference between AI Governance and AGI governance as related but distinct challenges. Following this, a brief overview of the literature on current proposals for International AGI Governance institutions is provided. Finally, key elements from these proposals are also sketched.. This section serves as the setup for a discussion of the Global AGI Agency Proposal in Part 3.

AI Governance & AGI Governance

AI Governance of Narrow AI Systems

As noted earlier, narrow AI systems have already had a significant impact on the global human cultural, social, and economic ecosystems. Digital computation, networked across the globe, laid the stage for these earlier AI systems to be integrated throughout our commercial, social, and political lives. The way in which narrow AI systems provided both important benefits and harms, at scale, points at how an AGI system might amplify even further these benefits and harms, alongside new, additional benefits and harms. As with our previous discussion of narrow AI, general AI, and AGI, we can imagine that impacts of narrow AI and emerging impacts of more general AI systems point at the types of consequences from an even more capable and more general AGI system.

AI Governance is a young field, but has mostly concerned itself with the consequences of narrow AI systems. This makes sense as, until recently, these were the only available AI systems. This research has mostly examined the impact of recommender algorithms in areas such as search, government services, ads, hiring, and social media. These recommender algorithms are still the form of AI that has arguably already had the greatest impact on humanity. The way in which these recommender narrow algorithms have shaped our consumption of information, and corresponding behavior is well documented. The widespread use of recommender algorithms began around 15-20 years ago, and today there are lots of governance efforts ongoing to attempt to set the appropriate guardrails, limits, and moderation practices.

AI Governance of General AI Systems

Now we are seeing a shift in focus for AI Governance as more general AI systems are being developed and deployed. From 2020-2022 it began to become clear to many observers that the new wave of advanced AI signaled a significant increase in the capabilities and generality of AI systems. The generality of these systems has prompted new concerns for AI governance. These new concerns have prompted a new wave of proposals for ensuring effective AI Governance.

A complete literature review will not be attempted here, as it has already been done by others. The most comprehensive of these reviews is Matthijs M. Maas and José Jaime Villalobos September 2023 Legal Priorities Project (now LawAI) report titled “International AI Institutions: A literature review of models, examples, and proposals.[14]” The authors note 7 “models” for International AI Institutions in particular, these types of models include: 1) Scientific consensus-

[14] https://matthijsmaas.com/uploads/Maas%20&%20Villalobos%20%282023%29_International%20AI%20Institutions%20%5BPublic%5D.pdf

building, 2) Political consensus-building and norm-setting, 3) Coordination of policy and regulation, 4) Enforcement of standards or restrictions, 5) Stabilization and emergency response, 6) International joint research, and 7) Distribution of benefits and access. This literature review highlights the various types of international institutions that could be created to help steer AI development and ensure it is governed well. However, as noted earlier, very few of these proposals directly engage with what would be needed once a single AGI has already been developed. A Global AGI Agency, containing the form of AGI that was described earlier, would need several of these models to be incorporated to ensure risks from AGI were minimized while benefits were also realized. At a minimum a Global AGI Agency would need processes for 1) Political consensus building and norm setting, 2) enforcement of standards or restrictions, 3) international joint research, and 4) distribution of benefits of access. In this way, this review of international AI institutions provides an excellent overview of what an international institution would need to be able to do to effectively govern a single AGI.

Another important and useful contribution in this direction is "International Institutions for Advanced AI[15]" by Lewis Ho and a host of coauthors posted to Arxiv in July of 2023. This paper identifies two broad institutional function categories for international governance of AI. These two broad categories are: 1) Science and Technology Research, Development and Diffusion, and 2) International Rulemaking and Enforcement. Within each of these broad categories the authors list several relevant and important functions. For the Science and Technology Research, Development and Diffusion category the authors note the following important functions: 1) Conduct or support AI safety research, 2) Build consensus on opportunities and risks, 3) Develop frontier AI, and 4) Distribute and enable access to cutting edge AI. For the International Rulemaking and Enforcement category the authors note the following important functions: 1) Set safety norms and standards, 2) Support implementation of standards, 3) Monitor compliance, and 4) Control AI inputs. While not explicitly focused on AGI, this paper begins to get at the heart of what would be needed for the world to effectively globally govern a single AGI.

Building from these insights Ho and colleagues organize these functions into four different plausible international institutions to govern advanced AI. These four institutions are: 1) An intergovernmental commission on Frontier AI, 2) An intergovernmental or multi-stakeholder Advanced AI Governance Organization, 3) An international public-private partnership Frontier AI Collaborative, and 4) An international AI safety project involving civil society and the private sector. The authors also note that the functions of these institutions could be merged together in various combinations. They state[16]:

[15] https://arxiv.org/pdf/2307.04699

[16] https://arxiv.org/pdf/2307.04699

"We can imagine institutions taking on the role of several of the models above. For example, the Commission on Frontier AI and the AI Safety Project make an obvious pairing: a Commission could scale up research functions to supplement the synthesis and consensus-building efforts, or a Project could conduct synthesis work in the course of its activities and gradually take on a consensus-establishing role. A Frontier AI Collaborative would also likely conduct safety research, and could easily absorb additional resourcing to become a world-leading Safety Project."

This paper identifies key elements that a Global AGI Agency would need, in a world with one AGI, however it stops short of providing a concrete plan for what such an agency would look like, what authorities it would have, who would be involved, and even identifying what pairing of these functions would be required. However, one final proposal does directly explore what institution specifically might be needed to effectively globally govern AGI. This is known as the MAGIC proposal.

AI Governance of AGI

The MAGIC proposal comes from the paper titled "Multinational AGI Consortium (MAGIC): A Proposal for International Coordination on AI[17]" written by Jason Hausenloy, Andrea Miotti, and Claire Dennis. This proposal tackles the governance challenge AGI presents head on, and seems to be unique in the AI Governance literature in doing so. The authors state that this proposal "has the following four core characteristics:

1. Exclusive: the world's only advanced AI facility, with a monopoly on the development of advanced AI models, and nonproliferation of AI models everywhere else.
2. Safety-focused: focused on the development of AI systems that are safe by design, including development of new architectures and ways to bound existing AI systems.
3. Secure: among the most highly secure facilities on Earth, with strict protocols for information security.
4. Collective: supported internationally, where the benefits of AI systems are distributed among all member countries."

As the authors state "MAGIC goes further than other proposals for international AI research institutions in its call for an immediate restriction of all external advanced AI development." To this end it is somewhat refreshing in the literature in that it attempts to fully wrestle with the existential and catastrophic risks presented by AGI.

The basic idea is that AI models, beyond some threshold of capabilities, should be exclusively developed by one institution. This institution would put safety and differential technological development as its main priority. It would ensure that AI models, research, and development are highly secure. And, it would be supported by an

[17] https://arxiv.org/pdf/2310.09217

international coalition of governments. While the authors do not sketch out a detailed institution, they do provide this key set of characteristics and discuss the broad strokes of what such an institution would need to contain.

In addition to the core characteristics noted above, the authors also argue that a safety focus may necessitate a move away from "black box" unitary agent frame to something more like the Open Agency AGI framework. They state that "Though we currently lack substantial empirical evidence supporting safer architectures, there appears to be a theoretical possibility of developing systems with clear boundaries and coordination to prevent alternative forms of development." And, in a footnote they suggest that "Another naively 'safe' system is the Open Agency model, which separates world models from data and planning and acting in real-time." They do note that these sorts of methods may involve a "safety tax" such that they may take more time to develop than simply throwing more compute at the problem, arguing that "All of these proposals share the property that these systems cannot compete with black box systems with increased amounts of compute." Nevertheless, a MAGIC-style proposal would allow for a deeper exploration of these approaches and a limiting of the sheer amount of compute that other actors would have access to.

Global AGI Agency Governance Components

One of the key goals of this report is to consider what might be globally needed in a world with AGI, and in particular a world with a specific form of AGI. As noted in Part 2, most of what is being considered in the space of AI Governance applies to pre-AGI systems. This too is useful, but given the pace of development of AI capabilities, more needs to be done to understand what would be required of a Global AGI Agency.

In September of 2024, the same month in which initial drafting for this report was completed, the UN multi-stakeholder High-level Advisory Body on Artificial Intelligence released their final report on Governing AI for Humanity[18]. This report, like many others, focuses primarily on what should be done now with respect to global AI governance. However, this body of global experts briefly notes that while they do not currently recommend establishing an international agency for AI, that if capabilities of AI systems continue to increase, such an agency may be necessary.

They state: *"eventually, some kind of mechanism at the global level might become essential to formalize red lines if regulation of AI needs to be enforceable. Such a mechanism might include formal CERN-like commitments for pooling resources for collaboration*

[18] https://www.un.org/sites/un2.un.org/files/governing_ai_for_humanity_final_report_en.pdf

on AI research and sharing of benefits as part of the bargain.

Given the speed, autonomy and opacity of AI systems, however, waiting for a threat to emerge may mean that any response will come too late. Continued scientific assessments and policy dialogue would ensure that the world is not surprised. Any decision to begin a formal process would, naturally, lie with Member States.

Possible thresholds for such a move could include the prospect of uncontrollable or uncontainable AI systems being developed, or the deployment of systems that are unable to be traced back to human, corporate or State actors. They could also include indications that AI systems exhibit qualities that suggest the emergence of "superintelligence", although this is not present in today's AI systems."

If an international AI agency were needed, say by the development of AGI systems then the UN report suggest that:

"The functions of a proposed international AI agency could draw on the experience of relevant agencies, such as IAEA, the Organisation for the Prohibition of Chemical Weapons, ICAO, IMO, CERN and the Biological Weapons Convention. They could include:

1. Developing and promulgating standards and norms for AI safety;
2. Monitoring AI systems that have the potential to threaten international peace and security, or cause grave breaches of human rights or international humanitarian law;
3. Receiving and investigating reports of incidents or misuses, and reporting on serious breaches;
4. Verifying compliance with international obligations;
5. Coordinating accountability, emergency responses and remedies for harm regarding AI safety incidents;
6. Promoting international cooperation for peaceful uses of AI."

Finally, the committee also points to four lessons learned from past global governance institutions. The lessons include:

1. "the development of a shared scientific and technical understanding of the problem is necessary to trigger a commonly accepted policy response"
2. "multi-stakeholder collaboration can deliver strong standards and promote quick responses"
3. "global coordination is often vital for monitoring and taking action in response to severe risks with the potential for widespread impact"
4. "it is important to create inclusive access to the resources needed for research and development, along with their benefits"

Now, given that the committee explicitly does not currently call for this an international AI agency unless AI systems are more dangerous or more capable, we can take their six functions and four lessons to be the most recent guideposts as to what many of the world's leading AI governance experts believe would be necessary for a Global AGI Agency. While an initial draft of this report was developed before the release of the UN Governing AI for Humanity, it complements these recommendations and provides some evidence of additional expert opinion on the mechanics and structure of a useful Global AGI Agency in a world with a single AGI.

One final, even more recent effort outside of the UN report that seeks to take not only AGI seriously, but the prospect of Artificial Superintelligence (ASI) is the report "A Narrow Path: How to Secure our Future"[19] by Andrea Miotti, Tolga Bilge, Dave Kasten, and James Newport. The focus of this report is preventing the development of ASI and building out a durable international governance system for advanced AI in the meantime. From a global governance lens this report calls for building "an international system that does not collapse over time." To do so, the authors call for the creation of three institutions: (1) International AI Safety Commission, (2) Global Unit for AI Research and Development, and (3) International AI Tribunal. These institutions, respectively, play the roles of: (1) central rule-setting body, (2) central, multilateral research lab, and (3) an independent judicial arm for resolving disputes.

Part 3: A Global AGI Agency Proposal

Introduction to the Proposal

Part 1: AGI Types and Part 2: AGI Governing Institutions have set the stage for the Global AGI Agency Proposal that will be described in Part 3 of this report. Part 1 described different possible types of AGI, and argued that the Open Agency AGI form allows for the many benefits of AGI systems while plausibly mitigating important risks from AGI such as power seeking and lack of transparency. The Global AGI Agency Proposal argues that if AGI is to be created, that the Open Agency AGI is the type of AGI that we should seek to develop. Part 2 summarized the state of research on AI Governance, pointed at the numerous proposals for governance of narrow AI systems, and briefly discussed proposals that directly addressed AGI governance and global governance of very advanced AI systems.

The proposals for global governance of AGI emphasize the wide array of both risks and opportunities that AGI presents to humanity. They also stress the importance of (1) effective global governance for the development of safe and transparent AGI and (2) the importance of the equitable distribution of the immense benefits an AGI system could bring to the world. To accomplish

[19] https://pdf.narrowpath.co/A_Narrow_Path.pdf

both of these goals, the Global AGI Agency is proposed.

At the highest level, the Global AGI Agency is a public-private coalition that is initially led by the major global AI state powers, which includes, ideally, the United States, China, the members of the European Union, and the United Kingdom as early coalition partners, with convening efforts supplemented by the United Nations. In addition to involvement by these major political powers, key private sector companies will also need to be deeply involved as partners in the coalition. If this agency were to be developed today, the key private sector companies that would need to be involved, ideally include, at a minimum, the US companies Alphabet, Meta, Microsoft, Amazon, OpenAI, and Anthropic, and the Chinese companies Baidu, Alibaba, iFlytek, Tencent, and SenseTime. Finally, in addition to deep involvement by nation states, supranational governing bodies like the EU and UN, and leading AI firms, professional technical organizations such as the IEEE should also be involved to provide broad guidance from its global technical expert members.

This multifaceted, multi-stakeholder approach is necessary given the global implications of AGI. No one country, region, or industry involvement is sufficient for coordinating AGI development, deployment, and governance. In what follows the report lays out the core elements of the Global AGI Agency proposal.

Core Elements of the Global AGI Agency

This section identifies and describes the core elements of the proposed Global AGI Agency. It will not provide a detailed blueprint for how to create such an agency, but rather will lay out the core elements of such an agency. The key elements that will be described at a high-level are: (1) The institutional framework of the agency, (2) the necessary points of global collaboration and integration, (3) the Open Agency AGI model itself, organized around principles of structured transparency, and (4) and the process of democratic accountability.

1. Institutional Framework

As described in the introduction to the proposal the Global AGI Agency will need to be multifaceted and include globally important stakeholders. This structure emphasizes that effective global governance must recognize the degree to which sovereignty over AI must consider the *de facto* nature of the hybridity of sovereignty in the world[20]. That is, nation states, global government bodies, large corporations, and large professional organizations all play prominent roles in exerting power in the world. The institutional framework includes: (1) joint support from the UN and IEEE, (2) coalition of governments and the private sector, and (3) a pooling together of the global resources that comprise the inputs to AGI.

[20] https://arxiv.org/pdf/2410.17481

The first key piece of the institutional framework for the Global AGI Agency is that it is supported jointly by the most prominent and relevant global institutions: the UN and the IEEE. The UN is the government body in the world with global reach. It has relevant experience and authority to help set appropriate governance standards. The recent UN report, cited earlier, "Governing AI for Humanity" makes this case in detail, and while the high-level committee did not recommend the UN currently establish an UN sponsored International AI Agency, it did detail the myriad ways in which the UN could help support such an agency. The report also noted the many limitations of an International AI Agency that would be placed within the structure of the UN. Given these strengths and limitations, the Global AGI Agency would benefit from being supported by the UN while remaining independent of its formal governance structure.

While the UN can play a key role in helping establish global governance standards and for helping create a network of states that is globally inclusive, the IEEE can assist with the needed technical expertise of its global membership of technological experts. The IEEE has a formal process for creating and adopting technical global standards with respect to new technological innovations. It also has legitimacy throughout the world in being the global expert that helps set these standards. Therefore the Global AGI Agency would benefit from also being supported by the IEEE. Taken together, the UN and IEEE should be integrated as key stakeholders and sponsors of the Global AGI Agency such that AGI development and deployment is overseen and guided by international best practices, technical standards, and respects global human rights.

In addition to the support of these relevant global institutions, the second key piece of the institutional framework for the Global AGI Agency is a coalition of governments and private sector companies. This coalition should begin with the most powerful and influential actors across nation states and large private companies. Ideally this coalition would eventually include every nation state, but it could begin as a coalition across the US and China, which are the two countries best situated to advance AI capabilities. These two countries are best situated to make these advancements because of the innovation of their leading companies. Thus, in addition to the support of the US and Chinese governments, the leading private companies within each country also need to be integrated into this coalition as well. As noted in the introduction, at the current moment in time these companies would include US companies Alphabet, Meta, Microsoft, Amazon, OpenAI, and Anthropic, and the Chinese companies Baidu, Alibaba, iFlytek, Tencent, and SenseTime. While the US and China do take different approaches to technological innovation, both countries have the tools and influence to require the participation of their leading AI companies, if desired. So the key piece here would be creating the appropriate incentives and structure such that both the US and China are willing to engage in this coalition. While the EU is not a major player in advancing AI capabilities, they have played a significant early role in convening nation states on issues of AI safety. It may be the case that both the EU and the UK could play significant political roles in helping create this coalition across the US and China. Finally, with these major players committed to this endeavor, it seems very likely that a host of other nation states would then be incentivized to join the coalition as well.

The third and final key piece of the institutional framework is global resource pooling. This piece is common to the MAGIC proposal and is often a component of CERN for AI proposals. The idea here is that the Global AGI Agency would pool together the relevant needed inputs to create an Open Agency AGI, rather than leaving its creation to a competitive effort across companies in an arms race for capabilities. The needed pool of resources would include leading global AI engineering talent, compute, energy, and data. The pooling together of these resources would provide the basic ingredients to create a single Open Agency AGI. Additionally, with these resources pooled together, there would only be a single coordinated, cooperative effort to develop AGI. While there are drawbacks to this approach, which will be discussed in Part 4, this approach allows for global transparency, inspection, and monitoring of the progress and capabilities of the AGI systems.

2. Global Collaboration and Integration

With each of the key elements of the institutional framework identified, the report will now briefly expand on the function of each of these elements to assist with global coordination and integration.

The key functions of the UN and IEEE match nicely to the proposed functions of an international AI agency identified in the UN Governing AI for Humanity report. These functions can be classified as:

1. Developing and promulgating global safety standards for the Open Agency AGI
2. Global monitoring of the Global AGI Agency
3. Investigating and auditing the agency
4. Verifying compliance of the Open Agency AGI with the global safety standards
5. Coordinating democratic accountability of the Global AGI Agency
6. Convening the relevant actors to ensure that the Open Agency AGI system distributes benefits equitably throughout the globe.

The UN and IEEE can partner across their comparative advantages to ensure global integration and collaboration across these various functions. The successful implementation of these functions will ensure that nation states and private sector contributions abide by these norms, standards, and protocols.

In addition to the global integrative and collaborative roles played by the UN and IEEE, collaboration across the initial nation state coalition is also vital for global collaboration and integration. It is at the nation state level where laws, rules, and norms can most readily be enforced for the behavior of the states themselves and the behavior of the private companies within those states. This is particularly important for the pooling of key resources to build, develop, and deploy the Open Agency AGI. And this is why the coalition needs to contain both the US and China. It is within these countries that the relevant access to engineering talent, computation, data, and electricity primarily reside. To be more specific, it is within the AI companies housed within these countries that these resources primarily reside, but it is the states themselves that have the capacity to ensure that partnerships with those companies ensure the acquisition of these resources. Furthermore, if the US and China, as the major AI developers in the world, can reach an agreement to pool their AI resources, it would set the stage for the inclusion of the rest of the world such that a truly global collaboration and integration can form. Again, it seems that the EU could play a significant role in helping to broker such a coalition, with support from both the UN and IEEE.

Finally, the contribution of the private sector should not be understated in the challenge of global coordination and integration. As discussed above and throughout, it is the private sector that contains most of the major inputs to developing, deploying, and overseeing the Open Agency AGI. Governments have simply not led in the creation of these systems. In particular, the key engineering talent, the immense amount of computation required, and the collection, cleaning, and preparation of massive amounts of data are all housed within private sector companies for which public-private partnerships are only just now beginning to be formed. Electricity to power these systems does have a stronger public-private legacy resulting from the long-standing oversight of utility companies, but the sheer amount of electricity required to power these systems will likely require significant private sector innovations and contributions.

It is unclear what the best form of private and public sector partnerships and collaborations are to ensure that AI companies are properly incentivized to participate, but both the Chinese and US governments have numerous tools at their disposal to aid in building out these collaborations[21] ranging from financial incentives to outright direct control of the management of the companies. However, as with the difficulty of securing a US-China coalition for the Global AGI Agency, there will also be significant challenges in developing and ensuring the needed private sector contributions are secured. This challenge is also discussed in Part 4.

3. Open Agency Model & Structured Transparency

So far in this section, we've only discussed aspects of the institution itself, with a particular focus on the

[21] https://www.convergenceanalysis.org/publications/soft-nationalization-how-the-us-government-will-control-ai-labs

institutional framework and how it enables global collaboration and integration. Missing is a discussion of the Open Agency AGI itself. As discussed earlier in this report, the Open Agency AGI has some particular advantages when compared to the Unitary Agent AGI. These advantages are particularly relevant to the Global AGI Agency proposal. An Open Agency AGI more readily allows for: (1) modular components, (2) bounded-role specific agents, and (3) structured transparency and control. Taken together these elements form a summative, distributed AGI model that allows for improved oversight, monitoring, and control. These components also somewhat buttress the concerns of centralization of power and power seeking behavior of an AGI system.

The Open Agency model as proposed by Drexler and currently pursued by multiple research programs, contrasts with the current Unitary Agent model in that it decomposes the AGI into numerous sub-systems whose inputs, throughputs, and outputs can be individually inspected. Each of these subsystems is a modular component of the overall system that is required for the AGI itself to complete tasks. This modular system allows for the AGI to be deconstructed, disassembled, and decommissioned as needed. It also allows for various technical experts to work on the system in parallel. Finally it allows for sub-systems to be added or subtracted based upon desired behavior, while also restricting access to key core components.

It is the modularity of the Open Agency AGI type that also allows for the creation of role-specific, bounded AI agents that are built on top of the Open Agency AGI. In this way the Open Agency AGI can be thought of as a complex, multifaceted "brain" and the bounded AI agents as the "actuators" that can engage in various behaviors that are limited by its access to various "regions" or "capabilities" of the brain. These bounded AI agents are thus given "structured access" to the more capable Open Agency AGI. This is not dissimilar from current structured access of LLMs through APIs.[22] This structured access has several useful features for accountability and transparency of the use of the Open Agency AGI system.

In the same vein as structured access, an Open Agency AGI can also be constructed in accordance with structured transparency. This model is applied to privacy in the paper "Beyond Privacy Trade-offs with Structured Transparency"[23] and generalized in a recent work from Drexler "Security without Dystopia: Structured transparency."[24] In Drexler's account structured transparency includes information flows and potential flow control mechanisms that allow for a system that has powerful pattern recognition (something like an LLM) housed with a secure information

[22] https://doi.org/10.1093/oxfordhb/9780197579329.013.39

[23] https://arxiv.org/pdf/2012.08347v1

[24] https://aiprospects.substack.com/p/security-without-dystopia-new-options

repository where much of a systems "intelligence" capabilities are housed, and where both the data repository and pattern recognition system can be considered their own subsystems within an Open Agency AGI. Access to these capacities is then controlled through flow control mechanisms such as redaction and anonymization, revocable permissions, rate control, time windows and query types. Each subsystem can be inspected by governance mechanisms and made available, in a focused, limited, and bounded way, to users of the system.

In this particular setup, the civil society partners that form the Global AGI Agency can oversee the governance mechanisms. The governance mechanisms have technical control over the intelligence capabilities and the flow control mechanisms. And the flow control mechanisms limit what can be inspected by the governance norms and standards.

Drexler describes it in "Security with Dystopia: Structured transparency" with the following example:

"Consider a potential transparency structure designed to enable detection and investigation of potential domestic security threats (perhaps plans for hijackings, bombs, or bioterrorism) while reliably precluding mass surveillance:

- AI systems operating inside an information-security boundary have access to rich information sources.
- AI-based pattern discovery can follow any clues, yet can report only specific threat-identifiers.
- Flow controls restrict human investigators to permissible, case-focused queries.
- Permissible queries are limited in number and scope, ensuring focused investigation rather than mass data collection.
- Substantial evidence of a serious threat can unlock access to broader information, a process similar to issuing a subpoena.
- Focused information is delivered to decision-makers for potential action."

And here, the same process could be applied to various bounded AI agents that individuals and organizations might want to deploy to accomplish specific tasks. This framework allows for restrictions for what the AI agents can access and how they can act. This framework also allows for continued research and development by the Global AGI Agency into improving data, AI-pattern discovery, and AI capabilities while providing meaningful control over access, deployment, and ensuring transparency and inspectability of the overall AGI system itself.

Taken together these elements of an Open Agency AGI create a summative, distributed AGI system that allows for responsible development, deployment, transparency, and control of the AGI system by the Global AGI Agency stakeholders.

4. Democratic Accountability and Access

Thus far the core elements of the proposal have provided a high-level overview of the institutional framework, how this framework enables global collaboration and integration across key powerful actors, and briefly discussed the benefits of the Open Agency AGI as the AGI type to be housed, created, and maintained by the Global AGI agency. The final core element is how this proposal also enables and fosters democratic accountability and widespread access that ensures both appropriate democratic feedback to the agency and equitable access to its capabilities.

The structured transparency architecture detailed by Drexler and discussed above highlights the pathway by which civil society should provide feedback to the governance mechanisms enabled by the agency. However these democratic mechanisms were left unspecified above. While the citizens of the US would have a form of democratic feedback through elected representatives, this is a very limited form of feedback. Democratic feedback can be enhanced through at least three mechanisms.

The first is to create public, globally accessible platforms for people across the world to both access and provide feedback on the services that are provided by the Global AGI Agency. While not discussed in detail above, the Global AGI Agency could enhance public services throughout the world by improving the overall capacity of governments to provide public goods such as transportation, healthcare, education, and social insurance. As part of these public services, citizens should be given a platform to provide direct feedback on the quality of these services such that these services can be continually improved through refinements to the Open Agency AGI.

Second public referendums should be provided for the global public to provide input on the governance standards, norms, and behaviors of the Global AGI Agency. This would help to ensure that the governance mechanisms themselves are globally inclusive and sensitive to regional and cultural context. Finally, there should be democratic input to the constitution of the AI agents that are deployed on behalf of both organizations and individual humans. Already the Collective Intelligence Project and Anthropic have experimented[25] with alignment assemblies and collectively-designed constitutions. Work of this sort would need to be expanded to set universal guardrails for the behavior of AI agents that can be accessed from the Open Agency AGI.

In addition to these democratic feedback mechanisms, democratic access of personalized AI agents is another important component of this proposal. As AI agents that are built on top of the Open Agency AGI are created, they will need to be both

[25] https://www.cip.org/blog/ccai

broadly accessible and "loyal[26]" to their respective users to enhance the equitable use of AGI and ensure that inequality is not further exacerbated. If these AI agents are both broadly accessible and loyal to individuals (within the bounds of a collective constitution) then they can unleash widespread welfare enhancements to humanity. This is a key feature of ensuring that the Global AGI Agency deploys AGI that is truly globally beneficial.

Proposal Summary

The Global AGI Agency Proposal has four core elements discussed above. The first is an institutional framework that includes: (1) joint support from both the UN and IEEE, (2) coalition of government and the private sector, and (3) a pooling together of the global resources that comprise the inputs to AGI. The second is global collaboration and integration which is ensured through: (1) a partnership by the UN and IEEE to help set global governance, technical standards, and convening, (2) an initial nation state collaboration across key AI powers such as the US and China agreeing to pool together their AI resources, (3) robust integration of leading AI companies. The third core element is the Open Agency AGI itself and the appropriate mechanisms to guide its development, deployment, and oversight, in which structured transparency plays a significant role. And the fourth core element is democratic accountability and access, which is enabled by: (1) democratic feedback platforms, and (2) public referendums and public participation in collectively-designed constitutions for AGI and AI agent behavior, and (3) widespread democratic access to AI agents.

While many of the details of this proposal need to be further elucidated, these core elements provide a high-level overview for the necessary components of a Global AGI Agency in a world where a single AGI system is created and deployed. This proposal calls for a global agency that is a partnership across global governance institutions such as the UN, global professional organizations such as the IEEE, leading nation state AI powers such as the US and China, and leading AI companies such as OpenAI and Anthropic. Furthermore it seeks to provide democratic input, access, and accountability. Finally it urges that the type of AGI that is created is an Open Agency AGI rather than a Unitary Agent AGI. While the exact details of the institutional design of the agency, the political process by which it would be created, and the technical design of the AGI system are left undeveloped here, it is hoped that these key elements highlight a guiding proposal for a world in which a single AGI system is developed that globally benefits humanity.

[26] https://doi.org/10.1093/oxfordhb/9780197579329.013.70

However, even at this high-level and with many details left undeveloped, the proposal itself does face important challenges which are explored in Part 4 below.

Part 4: Challenges

AGI will likely be the most powerful technology ever created by humanity. In the introduction, this report laid out how humanity has responded to other powerful technologies. In the mid-20th century, the two most powerful technologies humanity had created up to that point were birthed: nuclear weaponry and digital computers. As noted, these technologies have taken different routes in their influence on humanity. Nuclear weapons have remained in the hands of a few nation states, coupled with a host of global governance strategies to limit their immense power and coordinate their containment. On the other hand, digital computation has spread throughout the world like a wildfire. The internet globally connected these machines, and now digital computation, via access to the internet, is in the hands of two thirds of humanity. These competing approaches to how humanity has governed these powerful technologies provides competing analogies for how humanity should create and govern the development and use of AGI.

AGI combines the vast opportunity of global digital computation with the vast risks of nuclear weaponry. This suggests that neither analogy is perfect for how humanity should respond to the plausible creation of AGI. This report has attempted to lay out a high-level proposal for how humanity can learn from these competing examples. However, this proposal presents its own drawbacks and limitations for how the world might respond to the prospect of AGI. This section briefly discusses ten challenges for this proposal.

1. International Cooperation Challenges

The first significant challenge to this proposal is international cooperation. It was proposed that the US and China cooperate and enter into a coalition to pool their resources and share the power that advanced AI confers upon them with the world. Today's geopolitical tensions present a major hindrance to the possibility of such a coalition. US and China tensions are heightened, approaching something like a new cold war across these two global powers. This suggests that there will be a general reluctance of these two actors to share their AI resources. In addition to the current tensions, the US and China also present conflicting national values and priorities which further enhance the difficulty of cooperation and trust. Finally, even if cooperation can be secured across the US and China it is far from certain that global cooperation can be achieved.

2. Centralization of Power Risks

The proposal also presents centralization of power risks. While the proposal does include numerous elements to combat power centralization, it is difficult to dispute that a Global AGI Agency that oversees the development and deployment of AGI does not present significant risks in this direction. The creation of such an agency does present the creation of a single point of failure in AGI development and deployment. Additionally, the agency itself could become too powerful and too difficult to control. The agency itself could become compromised by bad actors and the AGI could be misused to nefarious ends. Furthermore, the “devil is in the details” for structuring the agency in which check and balances are maintained on a global scale across the numerous stakeholders that would be directly involved in this effort.

3. Innovation and Competition Concerns

Centralizing AGI research, development, and deployment risks reducing innovation and competition in the field of AI. If the agency is not properly designed, managed, and maintained, the bureaucracy of the agency itself could slow down progress towards beneficial AGI. This centralized approach also risks missing required diversity in research approaches and the exploration of unconventional development paths. Finally, the market rewards successful innovators generously, providing immense financial incentives for innovation. A coalition approach, led by the direction and oversight of governments, risks decreasing the financial incentives of innovators and thus making it difficult for a beneficial AGI system to be developed. The marketplace also weeds out bad ideas through competition, honing in on the good ideas. If development and deployment are shielding the Global AGI Agency from these forces, the resulting AGI system may simply struggle to be developed, denying humanity its immense benefits.

4. Private Sector Resistance

This proposal is also likely to receive significant private sector resistance, by US companies in particular. In the US context, technological companies are lightly regulated and generally left to create, experiment, and develop technological tools. Leading AI companies will be reluctant to surrender the competitive advantages that they have worked diligently to obtain. As noted with the previous challenge, it will also be difficult to properly incentivize private sector participation within the Global AGI Agency such that companies continue to innovate towards beneficial AGI. Many AI scientists and engineers may simply be unwilling to engage in this partnership and have a strong preference for completing their work in the private sector

5. Representation and Fairness Issues

While the proposal seeks an approach that ensures global representation and equitable distribution of benefits. This is an immense challenge. Even longstanding global bodies such as the UN and IEEE have perennially challenges with ensuring global representation. Additionally, the initial proposed coalition is among the major global power players both geopolitically and with respect to AI resources. This makes the global representation of the agency's governance a particular challenge. This approach risks marginalizing smaller or less technologically advanced nations. Furthermore, it is unclear how to balance the interests fairly across the directly involved stakeholders from governments, private sector, and civil society.

6. Complexity of Governance

Another challenge for this proposal is the complexity of the governance structure itself. The proposal calls for a diverse set of stakeholders across governments, private sector, and civil society. The interests of these various stakeholders diverge on many important questions. This divergence could result in deadlocks in the decision making process and present hard challenges for ensuring that the agency can respond quickly to the technological and governance challenges it encounters. This is in fact one common criticism of the UN, and the UN does not even attempt to directly incorporate large private companies within its formal decision making process. The agency will need intelligent and sophisticated mechanisms for resolving these governance challenges, and it is unclear what form these mechanisms should take.

7. Technical Challenges of the Open Agency AGI Model

The previous challenges have been centered on the governance challenges to this proposal, however technical challenges abound as well. This report argued that the Open Agency AGI model is a better approach for achieving AGI than the Unitary Agent AGI model. However, to date, most of the advances in AI capabilities can be found in increasing the size of systems that can more accurately be described as unitary models: large language and multi-modal models. While there are some efforts in developing alternative models for creating AGI, these models currently dominate. Given this, there are many unknowns about how to create an Open Agency AGI model that directly competes and surpasses the capabilities of Unitary Agent models. This presents plausible significant capabilities limitations for this type of AGI. It is also unclear how to develop true modularity and control of Open Agency models. These are open research questions for which it is difficult to know if they can be overcome to be the winning approach for the first true AGI system.

8. Democratic Accountability Limitations

One of the key elements of the Global AGI Agency proposal is democratic accountability and access, however securing meaningful and effective democratic accountability for a global agency is immensely challenging. To begin with, it is challenging for the general public to provide meaningful accountability of complex, cutting-edge technology developed by world-leading experts. In this direction it will be very difficult to appropriately educate the general global public on issues that AGI presents to them. Additionally, even in areas where the public's expertise is sufficient, integrating meaningful feedback into the development and deployment strategies of a technical agency is very challenging. Furthermore, the global nature of the agency and the technology make it such that global feedback is needed to provide true democratic accountability, and this is immensely challenging even in domains outside of cutting edge technology.

9. Security and Information Risks

One key, well-documented challenge, for current AI development is risk from cyberattacks. Current frontier AI companies are believed to have insufficient cybersecurity defenses.[27] These risks would be even more pronounced if a single agency was responsible for AGI development and deployment. This centralized approach ensures that the agency would be a target for espionage and cyberattacks. Given the diversity of stakeholders and individuals involved, there would be significant risks of potential leaks of key components of the Open Agency AGI system. There would be key, and currently unresolved, challenges in balancing desired transparency and needed security requirements in managing sensitive information. The proposed structured transparency approach seeks to address these challenges directly, but the challenges to security remain immense and difficult.

10. Adaptability and Future-Proofing Concerns

Finally, the last challenge for this proposal that will be discussed is adaptability and future-proofing. This proposal has been designed with a world in mind in which there is one AGI system. Thus it has been crafted to identify an ideal case in which the world has managed to create one, aligned, and controlled AGI. It has left unexplored, the challenges of a world that continues to evolve technologically. The Global AGI Agency, if successfully implemented, presents a plausible case in which a particular form of AGI can be globally governed to maximize the benefits and minimize the risks of a world in which the particular technology exists. However, it seems quite plausible that once this level of technology exists, further advancements in AI capacities will continue to develop. Once AGI exists, the prospect for Artificial Superintelligence (ASI) may also be imminent. It may also be impossible to contain the AGI capabilities for very long under the umbrella of one global agency. Furthermore, it is also difficult to prepare for unseen capabilities that even the Open Agency AGI may develop. These evolving technological

[27] https://situational-awareness.ai

challenges risk making the Global AGI Agency's structure obsolete. Finally, even without fast evolution of the single AGI's capacities, as AGI capabilities reverberate throughout the world, it is all but guaranteed to have dramatic impacts on the economic, cultural, and social features of our world. For the Global AGI Agency to remain relevant and achieve its goals of global cooperation and equitable global distribution of its benefits, the agency itself will need to have the capacity to evolve and adapt its structure, protocols, and decision making processes. This, too, is an immense challenge.

Part 5: Conclusion

This report offers a proposal for globally governing AGI in a world in which only one AGI system has been created. The proposal is titled the Global AGI Agency. The Global AGI Agency has four core elements. The first is the institutional framework. This framework is a joint UN-IEEE supported agency that is developed as a government-private sector coalition to pool global AI resources under the direction of one global agency. The proposal calls for global collaboration and integration across global institutions like the UN and IEEE, alongside an initial coalition of the US and China, leaders in AI development, and the private AI company powerhouses residing within those countries. Additionally, the proposal calls for a specific type of AGI that is described by the Open Agency AGI model which includes an AGI system that is modular, populated by bounded, role-specific AI agents, and organized around the principles of a structured transparency architecture. Finally, the proposal calls for democratic accountability and access through public feedback platforms, collectively designed AI-agent constitutions, and globally accessible, loyal AI agents.

This proposal has numerous strengths. First it seeks global cooperation and global integration through leveraging existing international bodies for their legitimacy and expertise. It encourages collaboration between the major AI powers of the US and China. It also integrates the capacities of the private sector by securing the involvement of the leading AI companies from within both the US and China. Second the proposal attempts to balance innovation of AGI while ensuring that the AGI is safe and controllable. The Open Agency model provides an architecture that is better suited to these goals than the Unitary Agent model. In addition to the Open Agency model the proposal calls for utilizing the structured transparency architecture for enabling inspection and technical and governance control of the AGI. The Open Agency model and the structured transparency architecture both create tools for facilitating improvement and feedback to the AGI and overall risk management of the system. Third, the proposal incorporates mechanisms for democratic accountability and global access through public feedback platforms, collectively designed AI-agent constitutions, and globally accessible, loyal AI agents. Finally, the proposal seeks to leverage the existing global governance institutions for their legitimacy and expertise. The proposal acknowledges and incorporates the role of governments, professional organizations, and private companies in the global governance system.

In addition to these strengths, the report also acknowledges the myriad challenges to the proposal. These challenges include:

1. International cooperation hurdles, especially US-China tensions
2. Risks of power centralization
3. Potential stifling of innovation and competition
4. Private sector resistance to resource pooling
5. Representation and fairness issues both for smaller nations and private companies
6. Complexity and weaknesses of multi-stakeholder, global governance institutions
7. Technical challenges of creating and maintaining the Open Agency AGI model
8. Limitations in achieving meaningful global democratic accountability
9. Security and information risks inherent in centralized development
10. Adaptability and future-proofing of the Global AGI Agency

In addition to these challenges, the proposal leaves many areas and questions for further research and refinement. The current proposal is a high-level, abstract collection of core needed elements for the development of a Global AGI Agency. Much more work is needed to describe in detail the needed institutional structure and strategies for incentivizing multi-stakeholder participation across both the powerful actors and public participation.

Beyond the specific strengths and challenges of this proposal, efforts have been made to collect and integrate the ideas from other international AI proposals. This previous literature has already identified both the necessity of and the opportunities and challenges of international AI governance. These proposals reviewed in Part 2 of this proposal provide a strong foundation for anyone who takes the prospect of the capabilities of AGI seriously and who is concerned with ensuring that the benefits of advanced AI are equitably realized and that the risks of these systems are minimized. While each of these proposals have their own strengths and limitations, this report has attempted to contribute to this literature by: 1) encouraging the creation of a specific type of AGI using the Open Agency model and the structured transparency architecture, 2) assuming a world in which AGI is developed, 3) expanding the stakeholders involved to reflect the role of global professional technical organizations such as the IEEE and direct involvement of the AI companies themselves, and 4) proposing specific mechanisms for democratic accountability and access.

GLOBAL
GOVERNANCE
INSTITUTE

Cours Saint Michel 30A
Brussels 1040, Belgium

info@globalgovernance.eu

www.globalgovernance.eu